DATA 612 Project Plan

Group Project

2022-10-14

Project Plan

Title of the project and type of analysis

Investigating two datasets on physicochemical properties and quality ratings of red and white wine samples by going through the entire data analysis process

Figure 1: .

What type of analysis?

We shall be performing descriptive Analysis, Exploratory Analysis (EDA) and Inferential Analysis (Research)

Names of the group members and their course (412 or 612)

1. Sunday Okechukwu (612)
2. Marti Sonko (612)
3. Josephus Nyumalin (612)

Description of the problem to be analyzed

In this project, We will go through the entire data analysis process to analyze a dataset on the chemical properties of wine and their associated quality ratings. Essentially, we will be analyzing two datasets, one, on red wine samples and the other on white wine samples from North of Portugal. Each wine sample comes with a quality ratings from from 1 to 10 and results from several physical chemical tests.

Our goal is to investigate how wine quality vary by Type, Alcohol content, Sugar, Acidity level and much more

Asking Questions

The first step of data analysis process is asking questions. A data analyst is someone who uses data to answer questions. Sometimes we ask questions first and get our data later and other times we get the data first and ask questions based on it. Here, we will practice asking questions with a real data set.

Along the way we shall answer the following questions:

1. Do wines with higher alcoholic content receive better ratings?
2. Do sweeter wines (more residual sugar) receive better ratings?
3. Is a certain type of wine (red or white) associated with higher quality?
4. What level of acidity (pH value) receives the highest average rating?
5. What chemical characteristics are most important in predicting the quality of wine?

Proposed data sources.

Wine Quality Data Set from UCI Machine Learning Lab.

There are two datasets that provide information on samples of red and white variants of the Portuguese "Vinho Verde" wine. Each sample of wine was rated for quality by wine experts and examined with physicochemical tests.

data is publicly available here and can be downloaded here red wine and white wine

Proposed approach for the Analysis

Data Analysis and Visualization with User Choices and Results

We organized the data analysis process into four steps shown below:

Step 1: Wrangle data

```
knitr::include_graphics("C:/Users/User/Documents/step_in_data_wrangling.png")
```

In this process we shall get the data we need in a form we can easily work with in three steps: gather, assess, clean. We will gather the data to help us answer the questions we posed earlier, assess the data to identify any problems in the data's quality or structure, and finally clean the data by modifying, replacing, or removing data to ensure that the dataset is of the highest quality and as well-structured as possible.

Step 2: Perform EDA (Exploratory Data Analysis)

Here, we shall explore and then augment the data to maximize the potential of our analyses, visualizations, and models. Exploring involves finding patterns in our data, visualizing relationships in the data, and building intuition about what we're working with. After exploring, we can do things like remove outliers and create better features from the data, also known as feature engineering.

3: Draw conclusions (or even make predictions)

This step is typically approached with machine learning which is beyond the scope of this course, however we shall focus on drawing conclusions with descriptive statistics and inferential statistics.

Step 4: Communicate results

Here, we shall attempt to justify and convey the meaning in the insights we have found, share what we have built, explain how we reached design decisions, and report how well it performs. There are many ways to communicate results: reports, slide decks, blog posts, emails, presentations, or even conversations. We shall use ggplot2 to communicate our findings in a visual display.

Statistical Modeling with User Choices and Results

Here we shall perform [illegible]st and hypothesis testing to compare the means of the two samples

Allocation of Responsibilities for the team

1. Data wrangling (Gathering, Assessing and Cleaning) : Sunday Okechukwu
2. Asking Questions and Building intuition about our data: Mart Sonko
3. Conducting Exploratory Data Analysis (EDA): Josephus Nyumalin:

Method and Timing for Collaboration

Method of collaboration shall [illegible]n-person meetings and virtual meetings to discuss the progress of the project at least four meetings in a week.

Project Steps/Schedule

Identify the start and end dates for each major phase:

- Gathering of Data - 10/19/2022 - 10/25/2022
- Data Preparation: Cleaning and tidying the data 10/26/2022 - 11/05/2022
- Conducting the analysis: EDA and Statistical - 11/06/2022 - 11/16/2022

- Producing the Group Report - 11/17/2022 - 11/26/2022
- Producing the Group Presentation Document - 11/27/2022 - 12/04/2022
- Rehearsing the Presentation - 12/05/2022 - 12/11/2022